REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT DATE	3. DATES COVERED *(From - To)*
25-09-2002	25-09-2002	27-02-2002 - 25-09-2002

4. TITLE AND SUBTITLE
New Event Detection

5a. CONTRACT NUMBER DAAH01-02-C-R034
5b. GRANT NUMBER
5c. PROGRAM ELEMENT NUMBER
5d. PROJECT NUMBER RTW11 02
5e. TASK NUMBER
5f. WORK UNIT NUMBER

6. AUTHOR(S)
Karen Lochbaum, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Knowledge Analysis Technologies
4940 Pearl East Circle, Suite 200
Boulder, CO 80301

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency
ATTN: ITO (Dr. Charles Wayne)
3701 North Fairfax Drive
Arlington, VA 22203-1714

10. SPONSOR/MONITOR'S ACRONYM(S) DARPA

11. SPONSORING/MONITORING AGENCY REPORT NUMBER

12. DISTRIBUTION AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.

20020927 165

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Intelligence organizations want to know when an unprecedented event or new information is reported. While there is good technology for searching, tracking, and filtering on known topics, current methods do poorly at detecting something new. The chief mechanism of search and topic tracking, spotting important words, is inappropriate ‹new stories are not ones with no important words. Because the degree of difference of new and old is different for different topics, uniform thresholds for overlap, as used in current filtering technologies, are also inappropriate. This project approaches the problem in three new ways. First, it applies Latent Semantic Analysis (LSA), a machine-learning technology that simulates human understanding of discourse. After automatic training on a large body of representative text, LSA accurately measures amount of meaning similarity between two passages using all the words in both. Texts with a few words in common are not judged similar if their meaning is different, but are, even if they use entirely different terminology, if their meaning is the same. Second, the system interacts with human users to adapt its criteria to their interests and the characteristics of the data. Third, it uses novel LSA-based storage and retrieval techniques to increase efficiency and capacity.

15. SUBJECT TERMS
First story detection, New event detection, Topic detection and tracking, Latent Semantic Analysis, Cross-Language information retrieval, Information retrieval, Information filtering

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
U	U	U	UU	2	Karen Lochbaum, Ph.D.

19b. TELEPONE NUMBER ***(Include area code)***
303-545-9092

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI-Std Z39-18

DARPA SB012-015 TITLE: New Event Detection
New Event Detection Final Report Phase I: September 24, 2002
Knowledge Analysis Technologies, LLC. Boulder, CO.

Table of Contents

Background and Introduction:

The problem. For many defense and intelligence applications, and for many commercial uses such as clipping services, it would be extremely helpful to find out quickly when new events occur or when new information appears about known events. Traditional information retrieval and search engine technology helps people locate information that they know to look for, but does not detect that something new has happened. DARPA-sponsored research in the area of Topic Detection and Tracking (TDT) has achieved good results on Tracking (finding more stories about a known event) but has not done so well on Detection, especially First Story Detection (finding the first reference to an unexpected event) and New Information Detection (identifying the new information in a series of stories about an event). Innovative approaches are required to solve the latter two problems. What are wanted are robust, accurate, general-purpose, language-independent approaches that minimize both false alarm and miss-rates, and have a parametric way to trade off those errors.
The opportunity: The textual information representation and retrieval method Latent Semantic Analysis (LSA, also known as Latent Semantic Indexing, LSI) offers a potentially valuable contribution to the solution of this problem. LSA represents the semantic content of complete paragraphs or larger texts, such as newspaper articles, intelligence reports, or message intercepts, as points in a high dimensional "semantic space." As shown by a variety of empirical tests, the similarity of the total content of two texts, as measured by their closeness in the space, accurately simulates the similarity of their meanings to humans. Texts that are highly similar in meaning, even if worded differently, have closely neighboring points, thus making it easy to determine that an item is not new. For example, in one application, LSA is used to detect plagiarism. If one of two documents has been modified by substituting synonymous words or phrases and changing the order of sentences, LSA nevertheless detects their very high similarity of semantic content, something that other information retrieval and detection systems cannot do nearly as well. On the other hand, almost any difference between two natural texts, no matter how small, will produce a different vector, making it possible to determine that something has been changed, if only by one word, the degree of difference being measured on a continuous scale. This capability is also better than that of most other

systems This is because it reduces false alarms due to wording changes that do not change meaning, and reduces misses that are often caused in other systems by their over-reliance on sets of important trigger words.

In other words, when appropriate, LSA will often represent two texts that share few or no words as highly similar in meaning, or two texts that share most of their words as not. Here are some dramatic examples: Similarity of meaning is measured on a scale of –1 to 1, with the mean for random passages ~ .08, sd .06.

Texts	Similarity of meaning
"measuring instrument for drive shaft velocity" - "gauge to determine propeller axle speed"	**.54**
"the radius of spheres" – "a circle's diameter"	**.55**
"the radius of spheres" - "the music of spheres"	**.01**

Results with other cases of the kind are not usually this strong, but on average they have been shown to be sufficiently good to yield a decided advantage over traditional information retrieval techniques, for example by ca. 30% in standard text information retrieval (Dumais, 1995). In this application the advantage promises to be much greater.

The trade-off between false alarms and hits is finely adjustable on a per-topic basis, with a potentially unlimited number of different topics at different thresholds, either by direct user selection or automatically by machine learning based on empirical choices made by analysts

Allan et al. (2000) "conclude that first story detection is either impossible or requires substantially different approaches." The approach we are following is substantially different in several respects. One is the removal of the term and document independence assumptions behind traditional IR methods; this assumption makes the capability illustrated in these examples impossible. Other innovations in our system are described below. These include automatic generation of brief summaries and new techniques for adaptive interactive training and use of the system by intelligence agents. We believe these can largely circumvent the "impossibility" that Allen shows is inherent in relying on near perfect automatic tracking as the fundamental approach.

The main reason that LSA can do different things from traditional IR is that it is pre-trained on a large body of representative text by a machine-learning process that infers the semantic similarity of every word to every other word. It combines representations of all the word meanings in a passage into representations of whole passage meanings in a manner that also approximates well the degree of sameness or difference to a human. [See Appendix for the mathematical basis and empirical verifications of LSA]. While this property of LSA is far from perfect, past results suggest that it is sufficiently effective to supply important functionality for a New Event detection system. Compared to other systems, LSA is relatively unlikely to think an old story new just because it is reported in different words, or a new story old just because it uses many of the same rare words. Testing how well this and other potential advantages of LSA are realized in practice was the accomplished overall goal of the Phase I research.

Overview of Phase I Accomplishments and Prototype Product.

Our prototype novel system for First Story and New Information detection operates by cumulatively processing and storing LSA-vector representations of whole news stories from an incoming stream and efficiently comparing each new item with all old items. While the working prototype has been built to perform all the essential functions of the envisaged system, it does not yet implement several of the useful features that these functionalities are designed to support, and contains only a small single-language corpus. However, newly developed techniques make it possible to scale LSA training to the corpus sizes that would be needed, and to rapidly update the system with new input. Moreover, the system is intrinsically applicable to any language; or any combination of languages. It can be automatically created for either single or cross-language use, requiring only a training corpus in each language (not specially parsed, aligned or enhanced.) Fast and economical parallel processing--in our work with Beowolf clusters--can give very rapid response times. As mentioned, the sensitivity of the system, the trade-off between miss and false alarm rate (signal/noise threshold), is continuously adjustable.

Here is an imaginary scenario of how the system might work in practice when fully developed. A human intelligence agent has identified a web site that carries discussions by members of a Neo-Nazi chat group in Holland, Germany and France that she wants to monitor. She copies its URL into her computer-agent. The next day, a box on her home page indicates a total of 50 new items from that source, 30 of them judged novel at default threshold, some of each in French, German, and English. She scans the 30 novel and potentially interesting messages using the automatically LSA-generated English summaries to decide if non-English notes are worth sending through machine translation. Each message is presented with bar indicators showing how unique it is relative to the other messages in the database. The most similar messages to each can be requested and returned in either similarity or recency order. She finds only a few that are of interest; most of the "novel" ones are personal notes of no interest. For each item she clicks [ignore ones like this],or [detect all information for ones like this]. For three, she asks for the earliest one above a particular similarity threshold similarity, and for one of these requests a search over the whole database for the five most similar items.

We should make it clear that this sketch is neither the same as the one presented in our original proposal, nor identical to the present version of the prototype system. Rather, it represents our current vision of the eventual system based on changes in design resulting from theoretical and empirical research in Phase I. It also needs to be understood that the LSA-based functionality proposed is not intended to be exclusive. We assume that other tools of other kinds, for example source-specific techniques that use clues such as standard titles, named entities, network links, and back-references, will be added in Phase II and III, and that the system may be combined with commercial high-powered, fast, but low accuracy search engines, or with new capabilities produced by other research groups.

What we have done is to construct a component for an eventual system that performs useful functions by exploiting LSA's advantages for the purpose. To convey the significance of this accomplishment, we list the special properties of the developed LSA-based features that we see as particularly important for new event detection:

- The fundamental engine of LSA is a semantic space that is derived in a completely automatic manner by machine-learning on large bodies of representative text. There is no expensive, time consuming hand coding of ontologies, dictionaries, grammars, or knowledge-bases. Given an appropriate training corpus, for example some millions of words from a similar source as that to be monitored, an LSA semantic space can be built, or updated, in a matter of hours.
- LSA is virtually language independent. Given an appropriate corpus in any language, in any machine readable orthography (e.g. Unicode), it can build a language-specific semantic space. Spaces have already been built for French, German, Spanish, Italian, Greek, Japanese in Kanji, Chinese, Arabic, Latvian, and for general and many discipline-specific genres of English.[1] This capability as applied to new event detection will be developed in Phase II.
- LSA can also build joint multiple-language semantic spaces by separate automatic training on each followed by a merging of the two spaces. (This capability was first demonstrated in an academic project supported by the CIA Office of Advanced Analytic Tools, and has been advanced by recent KAT R&D.[2]) There is some loss of accuracy in going across languages, but a useful level of performance is nevertheless obtained. Further research is needed, and would be performed in Phase II.
- LSA represents the entire semantic content of a text, not just that of selected or text-to-text matched keywords. The meaning of every word in both texts, and the complex manner of their combination, contributes to their computed similarity.
- LSA represents a continuous degree of semantic similarity between any two texts, which allows continuous adjustment of sensitivity.
- LSA's representation of text is at a level of meaning deeper than word overlap; texts can be estimated to be similar or different whether they use the same or different words. The similarities and differences so represented correspond well with what is similar and different for literate humans. One example of LSA's combined sensitivity to similarity despite vocabulary difference, and sensitivity to even small changes, comes from its use in detecting plagiarism in student essays. In a sample of 527 essays written on the same topic by students at a major Australian university, LSA was used to compare every essay with every other (138,601 comparisons). Twelve pairs of essays had similarities of .98 or .99, more than three standard deviations above the mean for the whole set. In several of these pairs a large number of words had been replaced by synonyms; in others, sentences had been rearranged. In this case it was easy to set a threshold that detected plagiarized essays and passed ones that were different. Remarkably, the two members of one of the plagiarized pairs had been read by the same

---

[1] Some languages, for example highly aglutinizing ones, may be less easily or well represented in LSA. More research is needed on this issue.

[2] The original technique for LSA-based cross-language retrieval gave the best results of any at the TREC conference (Rehder et al., 1998) at which it was introduced. Unfortunately, a presentation error gave rise to an audience misunderstanding that led some to conclude that it had instead done poorly.

professor within minutes of each other without detection. The technique has also detected copying in essay exams written by military officers.

- Qualitative results of Phase I research are consistent with our belief that it will be feasible for users to adaptively set thresholds that will separate old news from new news with useful accuracy. Systematic quantitative testing would be a central focus of the Phase II project.
- In LSA, the meaning of the combination of two texts is simply, empirically , and effectively computed by adding their separate representations. This supports a simple and elegant way to construct "prototypes" of old event classes against which to compare received items. This capability would be added in Phase II.
- Size constraints no longer a problem. Early applications of LSA, such as the IR experiments reported at several TREC conferences (Dumais, 1994, 1995, 1996, Rehder, et al, 1998), were constrained by limits on the size of matrices that could be handled by the Singular Value Decomposition packages and RAM capacities of the time. These constraints have been largely overcome by a new parallel SVD package released in 2001 by Michael Berry of the University of Tennessee, the availability of machines with orders of magnitude more RAM, and, especially, by new programs for which, in combination with parallel SVD, large arrays of independent processors can be used. For example, in a recent experiment, we created an LSA semantic space for a corpus equivalent in size to 200 years of AP newswire text. The processing took only 11 hours. In addition, the train-separately-then-merge technique described with reference to cross-language application can be applied to incrementally expand semantic spaces to unlimited sizes. While LSA may still not be the method of choice for whole-Internet search engines, for the data-stream monitoring task envisioned in the proposal, size constraints are no longer a problem.
- Frequent updating is no longer a problem. Similarly, in early applications, recomputing large semantic spaces took days or weeks, making it unsuitable for dynamic applications where new terms and topics arose with extremely high frequency (as is the case for Internet search-engines.) An updating technique called "folding-in", in which new documents were added on the basis of terms already present in the semantic space, provided an escape from the constraint, at the cost of a modest decrease in accuracy. However, current hardware and software make computation of even very large semantic spaces fast enough to be done many times a day. In addition, the train-separately then merge technique can be applied here as well.
- LSA is synergistically compatible with speech input. LSA representations of texts are based on the total combination of words. They are therefore relatively robust to automatic speech recognition errors at the word level. A few wrong words will not usually change the similarity between two documents by very much. It would very rarely change the selection of a nearest neighbor in semantic space because the effect would be in random directions in ~300 dimensions relative to the central meanings of the texts. These spaces are extremely sparse as well, so minor errors are unlikely to make a message similar to one it should not be similar to. New studies done in Phase I, as well as earlier pilot experiments, showed that even in the presence of substantial ASR errors, LSA was virtually unaffected. Developing speech input capability for new event detection would be conducted in Phase II.

Concept Demonstration of Working Prototype Interface and Functionality

A working prototype was developed to demonstrate features and functionality of a New Event Detector system. The prototype incorporates the 78,294 AP newswire articles from 1990; for any selected day in that year, it presents a list of the news articles that are considered most novel. Additional technical details are provided in the technical objectives and accomplishments section.

Figure 1 shows the main page of the New Event Detector. At the bottom left is a list of articles from February 8^{th} ranked by their "newness". In this case, newness is defined as being not similar to any article that has occurred previously in 1990. The initial assumption is that all articles from previous days have been read by the user and that she is now looking for novel events occurring on that day. Because of the trade-off between missing novel articles and false alarms on non-novel articles, the user can set a threshold indicating how strict she wants the system to be in defining what is considered a novel article (See the "Somewhat Novel" to "Very Novel" range selector in Figure 1). Those that fall above the threshold are indicated in red. Those below the threshold are indicated in black.

The interface presents a rank-ordered list of the novel articles for the day. Running the mouse over the article titles produces an LSA generated summary in the "summary" pane. When the user finds an article that is of interest to her, she can click on the title to bring up the complete article.

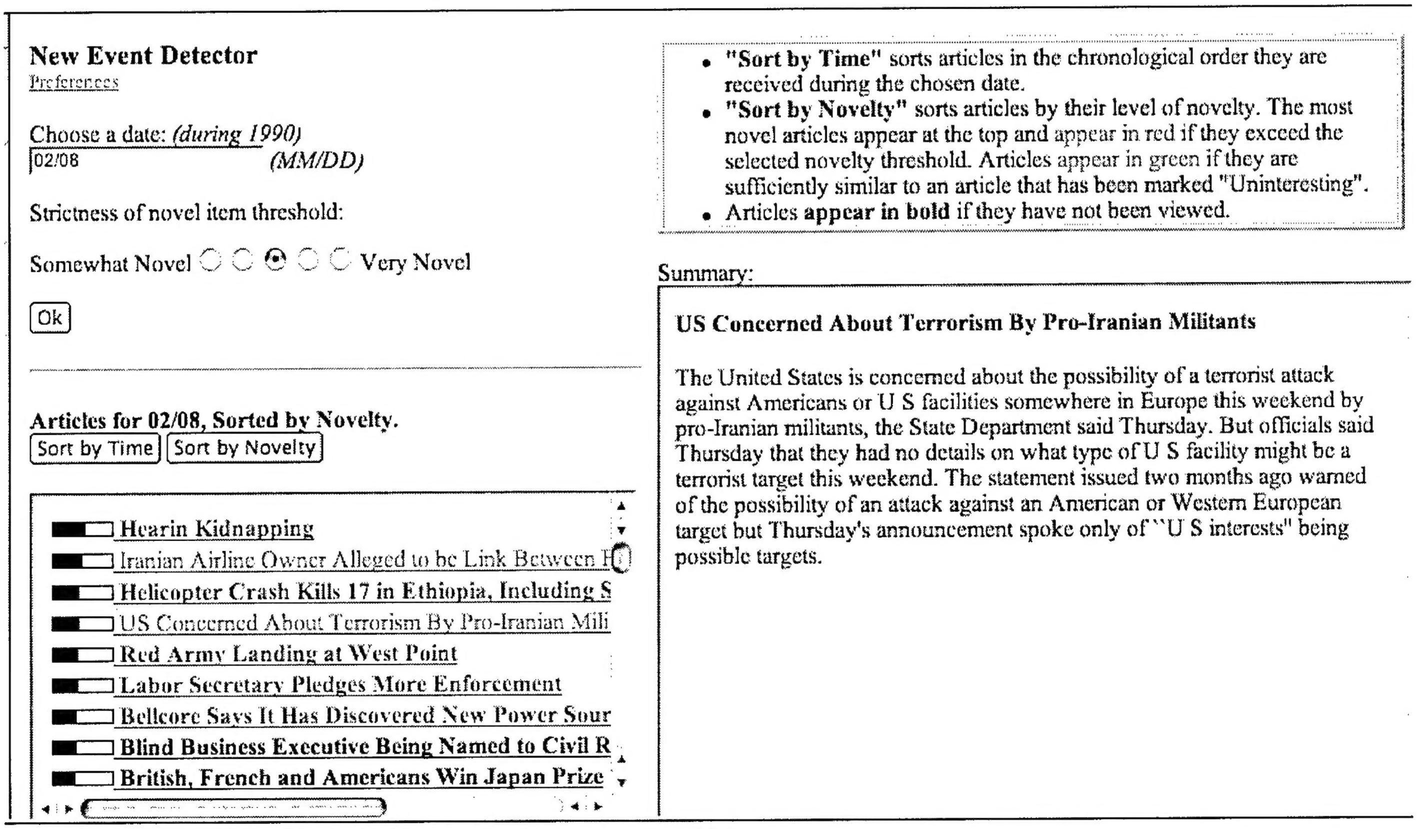

Figure 1. New Event Detector for February 8, 1990

In this scenario, the user clicks on the title "US Concerned about Terrorism by Pro-Iranian Militants." This action opens up a second window with the complete article (shown in Figure 2).

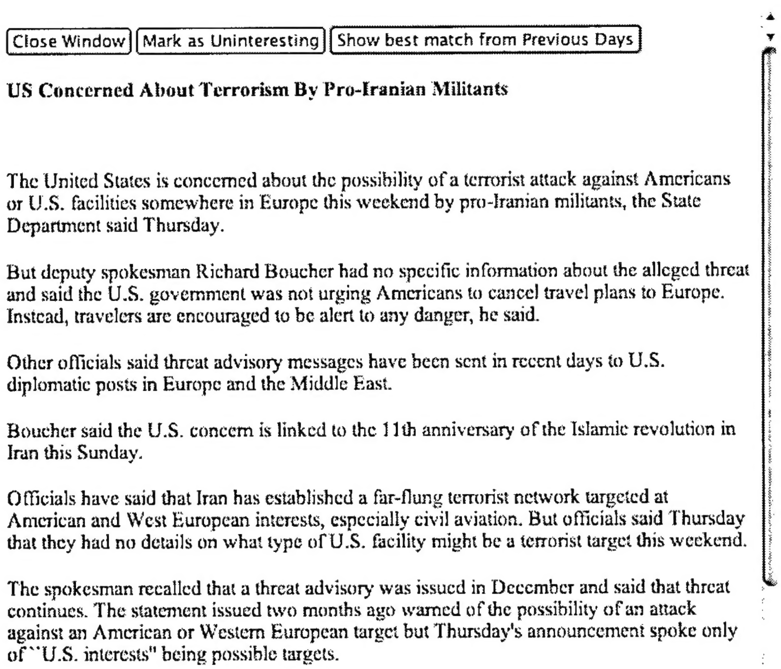

Figure 2. Selected article

Our user reads the article and wants to know if there have been any other articles related to this topic that have occurred previously. She clicks on the "Show best match from Previous Days" button and this opens a new window with the most similar article from the past (shown in Figure 3).

The window indicates that on February 7th, there was an article on the Pentagon being told to focus on regional threats in the Arabian Peninsula. While the article is similar in topic to the one identified on the 8th, the New Event Detector still judged that the original article from 02/08 was sufficiently "New" that it should be displayed (i.e. there were no previous articles in the user's profile specifically about Iranian terrorists.)

Close Window

Appeared 1990-02-07 18:46:00

Pentagon: Pay More Attention to Arabian Peninsula

By SUSANNE M. SCHAFER AP Military Writer

The Pentagon's military leadership has been told it should focus additional attention on possible regional threats to the Arabian Peninsula, a senior Defense Department official said Wednesday.

The guidelines are in the Defense Planning Guidance, the long-range planning instructions issued by Defense Secretary Dick Cheney to help the services prepare their budgets for 1992-'97.

The document was due last fall, but its release was delayed so planners could assess the changes that are sweeping the Soviet Union and Eastern Europe.

The senior Defense Department official briefed a number of reporters about the classified document after reports of its contents surfaced in the media.

Asked which regional threats are possible against the nations on the Arabian Peninsula, the official said the document does not cite nations by name, but he went on to say that "Iraq and Iran have significant armed forces."

Figure 3. Article that best matches from previous days

After reading any article, the user has the ability to mark that article as "uninteresting". If we assume that our user now feels fully informed about the developments of Iranian terrorists and no longer wants to see articles on that topic, she can click on the "Mark as Uninteresting" button in Figure 2. This updates her user profile indicating that new articles related to Iranian terrorists would not be considered novel. (It should be noted that this can also be done in the inverse for filtering, in which a user can indicate an article as "interesting" and any new articles on that topic would be highly ranked.)

If our user logs in the next day (February 9th), she would see what is shown in Figure 4. Ranked at the bottom of her list of articles (i.e. the least "novel" articles) is an article titled "Iran Says U.S. charge of planned attack baseless". This article is marked in green indicating that it was sufficiently similar to the "uninteresting" article she had added to her profile and therefore the new article was not considered novel. Thus, she can iteratively update her profile as well as adjust current thresholds as to what topics at what degree of novelty she would like to track.

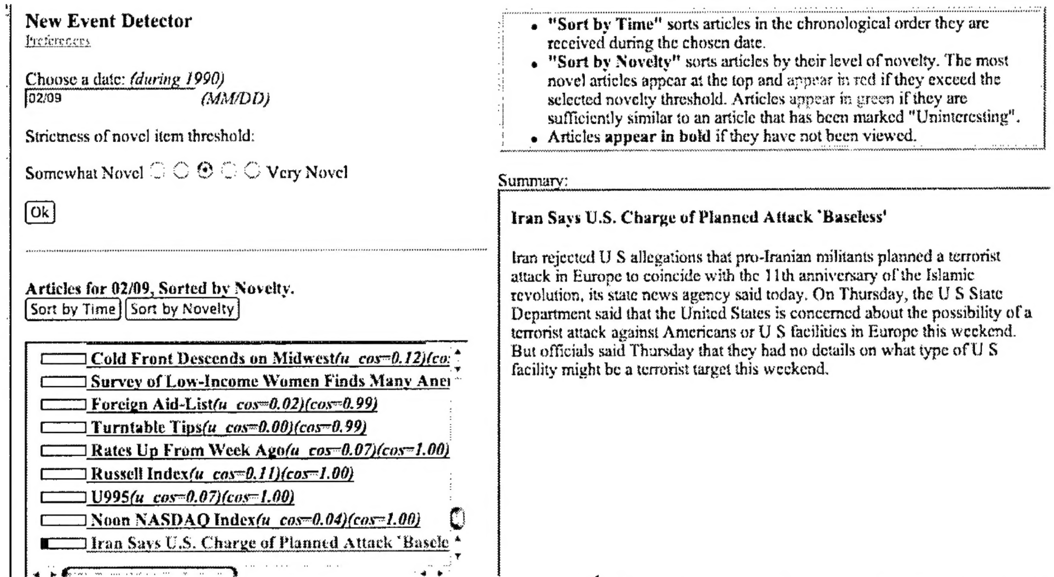

Figure 4. New Event Detector for February 9th showing the an article that is similar to one in the profile that has been marked as "uninteresting".

Specific Phase I Technical Objectives and Accomplishments.

Approach.

The R&D approach for this project was quite straightforward. It consisted of modifications and refinements of our previously created information finding tools, their integration into a novel system with new functionality, and evaluation of its feasibility and promise preparatory to implementation of a full scale prototype system in Phase II. The Phase I accomplishments are listed in correspondence to the Phase I technical Objectives described in the Phase I proposal.

1. Objective:1(a). Create an LSA semantic space of a large newswire or newspaper publication over a period of two or more years for research and testing. For the Phase I research, we created an LSA semantic space of all the AP articles from 1990 (108,892 terms, 78,294 documents). Objective 1(b): Expand this corpus and explore use of the TREC-TDT and TREC-Novelty datasets as well. After some exploration and experimentation, we concluded that the new TREC TDT and Novelty databases would neither be ready early enough for use in the project nor suitably constructed for our needs because they eventually assumed different problems than those in the objectives of this project.

Objectives 2 and 3. Develop a radial basis (RBF) function neural-net training. and an automatic prototype formation functionality based thereon. After theoretical and empirical research we determined that this proposed technique was no longer optimal. Recent advances in the size of corpora that can be handled by LSA and in the speed with which novelty measures can be computed against large databases obviated the need for pre-clustering, permitting us to use a more accurate and adaptable exhaustive comparison approach. The new novelty evaluation algorithm computes LSA cosines with all past articles, using the feedback from the user to adapt its rankings of displayed documents according to the user's specific interests as represented by similarity to past documents selected as interesting or not, ordered by their overall novelty, and calling attention to those above a dynamically settable threshold. A Phase II effort would improve efficiency for extremely large collections by inserting a non-LSA pre-filter to shortcut search of uninteresting documents.

Objective 4. Develop a standing filter component as it will be applied to a news feed. This was accomplished and tested with favorable results in prototype and favorable projections for a completed system. (See Objectives 2 and 3 above for description.)

Objective 5. Determine by experiment and analysis the effects of variation in parameters for thresholds and RBF on miss and false alarm rates. As stated under Objectives 2 and 3 this objective was obviated by the altered design, however considerable exploration of the best ways to set and display thresholds and feedback presentation of degree of novelty was conducted, resulting in the design illustrated in the working version 0 prototype.

Objective 6. Conduct initial experiments with users to judge the effectiveness of interactive system training and use. This was accomplished to a useful degree through iterative trial use by KAT employees. The current user-system interaction protocol appears both at least minimally adequate and suggestive of considerable improvement with more systematic usability engineering in Phase II.

Objective 7. Create a web-based exploratory prototype sufficient for examination and trial use, and to serve as a test-bed for further research and development. This was accomplished as proposed, and became a central focus in later periods of the Phase I R&D, involving a large number of revisions and evaluations, and resulting in what we believe will be, after additional work in Phase II, a better interactive environment for analysts than originally anticipated.
Objective 8. Determine the feasibility and estimate quality of performance of a fully developed system. Feasibility was clearly established, eventual quality and speed of performance as extrapolated from that of the prototype will be adequate, and unanticipated advances in user interface and functionality were made.

Additional accomplishment not originally proposed. We perfected and added a summarization feature that, using LSA, presents a three-sentence summary of any chosen article to make rapid navigation of the database easier and more effective. We performed testing to verify that this summarization approach produced more effective summaries than other approaches (e.g., MS-Word summaries, initial sentences or first paragraphs of news articles)

Following is a summary of effort and results keyed to the proposed work plan.

Work Plan.

Task no.	**Description.**
Task 0.	Complete R&D plan, staffing, schedule, COR kickoff, first progress report.**Done.**
Task 1.	Obtain, expand, and/or create corpora for an LSA semantic space of a large newspaper publication over a period of two or more years. **Done**. (but only one year, 1990, used in prototype.)
Task 2.	Develop radial basis (RBF) function neural-net training component. **Explored and abandoned in favor of a better approach, which was completed.**
Task 3.	Develop the automatic prototype formation functionality. **Explored and abandoned in favor of better approach, which was completed.**
Task 4.	Develop the standing filter component. **Done in a new and presumably superior manner**. **Added auto-summary feature to aid user navigation and training of system**
Task 5.	Determine thresholds. **Explored an adaptive approach which was developed and implemented.**
Task 6	Tests of user-interactive system training. **Formative tests done.**
Task 7.	Create a web-based exploratory prototype. **Done; more effort and accomplishment than proposed**
Task 8.	Estimate feasibility and performance quality of full

	system, prepare final report and Phase II proposal. **Feasibility well established.**

Planned Phase II follow-up. Phase II would develop a fully functional prototype operating over multiple sources of input, including chosen web-sites and real or emulated intelligence feeds, in more than one language of interest to the sponsor. It would also refine and improve the component detection algorithms, and add other useful features such as ways to classify and characterize new events by type and by what is new about them. One important improvement over the prototype will be to make the novelty threshold (how dissimilar to it an event should be to it to be called novel) differentially settable by adjustments applied permanently to single documents by the user. By this means the system will gradually be taught by each analyst to accept news of their interest more leniently and reject other kinds more strictly. We would also improve efficiency for extremely large collections by inserting a non-LSA pre-filter to shortcut search of uninteresting items. Additional advances will be implemented by augmenting the similarity measuring technology with statistical learning and text categorization models that have recently been developed in our own and other research efforts. These add capabilities to include syntactic components and better ways to characterize meaningful topic classes that are and are not of interest. These technologies include the SP model of our consultant, Prof. Simon Dennis, Support Vector Machines, and the Topics model recently invented at Stanford University by Prof. Mark Steyvers, now at UC Irvine, and Tom Griffiths, now at MIT.

Commercialization Plans.

Government and military intelligence agencies have a pressing need for this kind of system. Because making it ready for other kinds of textual information streams will be only a matter of automatic creation of a new LSA semantic space and interactive training by its actual users, re-purposing for other applications will be relatively easy. Among the commercially attractive applications are news organizations, Patent Office and private patent and product search companies, corporate intelligence, and clipping services. KAT would seek commercialization by sale to DOD and Federal intelligence agencies, direct sale of Internet based customized news and clipping services, and channel marketing through existing topical news collection and dispersal companies. We will work through the Navy CAP program, as we have in the past, and through our existing business and research contacts in the CIA, the Air Force, SAIC, and others, to identify organizations that will procure this technology.

Appendix A. Technical background: Latent Semantic

LSA is a machine-learning technology for simulating human understanding of the meaning of words and text. It uses a fully automatic mathematical/statistical technique to extract and infer meaning relations from the contextual usage of words in large collections of natural discourse. It is not a traditional natural language processing or artificial intelligence program; it uses no humanly constructed ontologies, dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies. LSA takes as input only large quantities of raw text parsed into words defined as unique character strings and separated into meaningful passages such as sentences or paragraphs. Although it is based on the statistics of how words are used in ordinary language, its analysis is much deeper and more powerful than frequency, co-occurrence, or keyword counting and matching techniques. In well-demonstrated ways, LSA simulates practical aspects of human meaning to a highly useful level of approximation. It can perform full-scale, significant tasks that, when performed by a human, depend on understanding the meaning of textual language. Its earliest applications were to textual information retrieval, where, other things equal, it out-performed most other systems, including ones developed with orders of magnitude more effort (Dumais, 1997.) Its advantages for the present application are that it is very easy to use and that it correctly estimates the degree of meaning similarity between two documents whether or not they are phrased in or indexed by the same words.

How LSA works. The problem of language learning (for either a machine or a human) can be represented as follows. The meaning of a passage (psg) is some function of the meaning of its words, plus the meaning of some non-verbal context.

m(psgi) f{mwdi1, mwdi2, ... , mwdin, mXi}].

The language learner's problem is to solve a an enormous system of such simultaneous equations for the meanings of all the words in a language or sub-language, and thus also the meaning of any passage. To make approximate solution feasible, LSA makes three simplifying assumptions: non-verbal context is ignored; only textually representable language is used as data, and the word-meaning combination function is addition.

m(psgi) ~ f{mwdi1+mwdi2 + ... + mwdin}

A large corpus of text, as similar as possible to the sources from which the humans whose judgment is to be simulated would have acquired the knowledge to be simulated, is divided into meaningful passages, such as paragraphs, which are then represented as equations. An approximate solution to the system of equations is found by the matrix algebraic technique of Singular Value Decomposition. (For details see Berry 1992, Berry, Dumais and O'Brien, 1995, Deerwester et al., 1990, or Landauer and Dumais, 1994, 1996, 1997). Each word in the corpus, and any passage, is represented as a high (typically around 300) dimensional vector. The non-monotonic optimal number of dimensions is important. Dimension reduction constitutes an inductive step by which words are represented by values on a smaller set of abstract features rather than their raw pattern of observed occurrences. One important result of this is that the ~98% of words that never appear in the same document, such as different terms for the same thing, get appropriately represented. Relations between meanings are computed as cosines between reduced-dimensional vectors.

This model yields good simulation of human verbal meaning across a wide spectrum of verbal phenomena and test applications: (1) correct query-document topic

similarity judgments, even when there are no literal words in common between query and document (Berry & Dumais, 1995; Caid, Dumais & Gallant, 1995; Deerwester et al., 1990; Dumais, 1991, 1993, 1994, 1995, 1996; Dumais & Nielson, 1992; Dumais & Schmitt, 1991, Foltz, 1990; Foltz & Dumais, 1992;); (2) correctly mimicking human word-word semantic relations and category membership judgments (Landauer, Foltz and Laham, 1998), (3) correct choices on vocabulary, and, after training on a textbook, subject-matter multiple choice tests (Landauer, Foltz and Laham, 1998), (4) accurate measurement of conceptual coherence of text and resulting comprehensibility (Foltz, Kintsch and Landauer. 1998), (5) correctly predicting word-word and passage-word priming of word recognition in psycholinguistic experiments (Landauer and Dumais, 1997), (6) accurate prediction of expert human holistic ratings and matching of the conceptual content of student essays and textbook sections (Landauer, Foltz and Laham, 1998), (7) optimal matching of instructional texts to learner knowledge as displayed in essays (Wolfe et al., 1998), (8) correctly mimicking synonym, antonym, singular-plural, past and present tense, and compound-component word relations, (Landauer, Foltz and Laham, 1998), (9) representing word ambiguity and polysemy, the possession of two or more distinct senses or meanings by the same word, (10)correctly mimicking semantic categorical clusterings of words, including ones found in certain neuropsychological deficits (Laham, 2000), (11) providing significant improvement for language modeling in Automatic Speech Recognition (Coccaro and Jurafsky, 1998), (12) matching textual personnel work histories to discursive job and task descriptions (Laham and Bennett, 2001), (13) estimating conceptual overlap among large numbers of training courses by analysis of test contents (Landauer, Foltz and Laham, 1998), (14) accurately simulating the growth of human vocabulary during K-12 school years (Landauer and Dumais, 1997.)

References

Allan, J., Lavrenko, V. and Jin, H. (2000). "First Story Detection in TDT Is Hard," in the *Proceedings of the Ninth International Conference on Information and Knowledge Management (CIKM)*: 374-381.

Berry, M. W. (1992). "Large scale singular value computations." *International Journal of Supercomputer Application 6(1)*: 13-49.

Berry, M. W., Dumais, S. T., and O'Brien, G. W. (1995). "Using linear algebra for intelligent information retrieval." *SIAM Review, 37(4)*: *573*-595.

Caid, W. R., Dumais, S. T. and Gallant, S. I. (1995), "Learned vector space models for information retrieval." *Information Processing and Management, 31(3)*: 419-429.

Coccaro, N. and Jurafsky, D. (1998). "Proceedings of the International Conference on Spoken Language Processing." ISSLP-98.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.A. (1990). "Indexing By Latent Semantic Analysis." *Journal of the American Society for Information Science 41(6)*: 391-407.

Dumais, S. T. (1991). "Improving the retrieval of information from external sources." *Behavior Research Methods, Instruments and Computers, 23(2)*: 229-236

Dumais, S. T. (1993). "LSI meets TREC: A status report." In: D. Harman (Ed.), *The First Text REtrieval Conference (TREC1), National Institute of Standards and Technology Special Publication 500-207*: 137-152.

Dumais, S. T. (1994). "Latent Semantic Indexing (LSI) and TREC-2." In: D. Harman (Ed.), *The Second Text REtrieval Conference (TREC2), National Institute of Standards and Technology Special Publication 500-215*: 105-116.

Dumais, S. T. (1995). "Using LSI for information filtering: TREC-3 experiments." In D. Harman (Ed.), *The Third Text REtrieval Conference (TREC3) National Institute of Standards and Technology Special Publication.*

Dumais, S. T. (1996). "Combining evidence for effective information filtering." In *AAAI Spring Symposium on Machine Learning and Information Retrieval, Tech Report SS-96-07*, AAAI Press, March 1996.

Dumais, S. T. (1997). "Using LSI for Information Retrieval Information Filtering, and Other Things." Talk at Cognitive Technology Workshop, April 4-5, 1997.

Dumais, S. T., Furnas, G. W., Landauer, T. K. and Deerwester, S. (1988), "Using latent semantic analysis to improve information retrieval." In *Proceedings of CHI'88: Conference on Human Factors in Computing*, New York: ACM, 281-285.

Dumais, S. T., and Landauer, T. K. (1984). "Describing categories of objects for menu retrieval systems." *Behavior Research Methods, Instruments and Computers 16(2)*: 242-248.

Dumais, S. T. and Nielsen, J. (1992). "Automating the assignment of submitted manuscripts to reviewers." In N. Belkin, P. Ingwersen, and A. M. Pejtersen (Eds.), *SIGIR'92: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM Press: 233-244.

Dumais, S. T. and Schmitt, D. G. (1991). "Iterative searching in an online database." In *Proceedings of Human Factors Society 35th Annual Meeting*: 398-402.

Foltz, P. W. (1990) "Using Latent Semantic Indexing for Information Filtering". In R. B. Allen (Ed.) Proceedings of the Conference on Office Information Systems, Cambridge, MA, 40-47.

Foltz, P. W. and Dumais, S. T. (1992). "Personalized information delivery: An analysis of information filtering methods." *Communications of the ACM, 35(12)*: 51-60.

Foltz, P. W., Kintsch, W., Landauer, T. K. (1998). "The measurement of textual coherence with Latent Semantic Analysis." *Discourse Processes 25(2&3)*: 285-307.

Laham, D. (2000), *Automated content assessment of text using Latent Semantic Analysis to simulate human cognition*, Ph.D. Dissertation, University of Colorado, Boulder.

Laham, D., Bennett, W., and Landauer, T. K. (2001). An LSA-based software tool for matching jobs, people, and instruction. *Interactive Learning Environments.*

Landauer, T. K. and Dumais, S. T. (1994). Latent semantic analysis and the measurement of knowledge. In R. M. Kaplan and J. C. Burstein (Eds.), *Educational Testing Service Conference on Natural Language Processing Techniques and Technology in Assessment and Education.* Princeton, NJ, Educational Testing Service: 127-141.

Landauer, T. K. and Dumais, S. T. (1996). "How come you know so much? From practical problem to theory." In D. Hermann, C. McEvoy, M. Johnson and P. Hertel. (Eds.), *Memory in context.* Hillsdale, N.J, Lawrence Erlbaum Associates.

Landauer, T. K. Dumais, S. T. (1997). "A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge." *Psychological Review 104*: 211-240.

Landauer, T. K., Foltz, P. W., Laham, D. (1998). "An introduction to Latent Semantic Analysis." *Discourse Processes 25(2&3)*: 259-284.

Rehder, B., Littman, M. L., Dumais, S. T. and Landauer, T. K. (1997). "Automatic 3-language cross-language information retrieval with latent semantic indexing." *The Sixth Text Retrieval Conference Notebook Papers (TREC6)*, National Institute of Standards and Technology Special Publication.

Wolfe, M.B., Schreiner, M.E., Rehder, B., Laham, D., Foltz, P.W., Kintsch, W., Landauer, T.K. (1998). "Learning from text: Matching readers and text by Latent Semantic Analysis." *Discourse Processes 25(2&3)*: 309-336.